Improving Product Classification Using Images

Anitha Kannan
Microsoft Research Search Labs
Mountain View, USA
ankannan@microsoft.com

Partha Pratim Talukdar*
Carnegie Mellon University
Pittsburgh, USA
partha.talukdar@cs.cmu.edu

Nikhil Rasiwasia*
U. of California, San Diego
San Diego, USA
nikux@ucsd.edu

Qifa Ke
Microsoft Research
Mountain View, USA
qke@microsoft.com

***Abstract*—Product classification in Commerce search (*e.g.,* Google Product Search, Bing Shopping) involves associating categories to offers of products from a large number of merchants. The categorized offers are used in many tasks including product taxonomy browsing and matching merchant offers to products in the catalog. Hence, learning a product classifier with high precision and recall is of fundamental importance in order to provide high quality shopping experience.**

A product offer typically consists of a short textual description and an image depicting the product. Traditional approaches to this classification task is to learn a classifier using only the textual descriptions of the products. In this paper, we show that the use of images, a weaker signal in our setting, in conjunction with the textual descriptions, a more discriminative signal, can considerably improve the precision of the classification task, irrespective of the type of classifier being used. We present a novel classification approach, Confusion Driven Probabilistic Fusion++ (CDPF++), that is cognizant of the disparity in the discriminative power of different types of signals and hence makes use of the confusion matrix of dominant signal (text in our setting) to prudently leverage the weaker signal (image), for an improved performance. Our evaluation performed on data from a major Commerce search engine's catalog shows a 12% (absolute) improvement in precision at 100% coverage, and a 16% (absolute) improvement in recall at 90% precision compared to classifiers that only use textual description of products. In addition, CDPF++ also provides a more accurate classifier based only on the dominant signal (text) that can be used in situations in which only the dominant signal is available during application time.

***Keywords*-product classification, e-commerce, text, image**

I. Introduction

Online shopping has revolutionized the way we shop. US online retail spending surpassed $155B in 2009, and is expected to surpass $248B in 2014 [1]. A recent Nielsen study reports that over 80% of American Internet users will make an online purchase in the next six months [2]. The last decade has witnessed a surge of commercial portals (such as Amazon), comparison shopping sites (such as PriceGrabber and NextTag), and commerce search verticals (such as Google Product Search and Bing Shopping).

In the commerce search setting, description of products, referred to as ***offers*** in this paper, are available from a variety of sources, including product manufacturers (*e.g.,* Sony, Canon, Samsung, *etc.*), data providers (*e.g.,*Etilize), and merchants (*e.g.,* Target, Walmart, BestBuy, *etc.*). Offers usually consist of a brief textual description of the product (often, the title of the product), and sometimes its corresponding image. While some offers are categorized by offer creators, majority of them are uncategorized.

Since online shopping sites target a rich and diverse set of products, a key to their success is to provide the ability for users to browse offers of products organized according to the product taxonomy used by the commerce search engine. In addition, such categorization becomes central for the task of offer to product matching where the goal is to identify the product that is in correspondence with the offer [11]. Thus, automatic classification of offers under the taxonomy is of fundamental importance.

Existing approaches to offer classification rely purely on the textual description of the offers [3], [19]. However, classifiers that rely exclusively on text suffer from the following problems:

1) **Overlapping text across categories:** Many categories in the taxonomy have products that are interrelated, and thus the textual descriptions of their products overlap in vocabulary usage. For example, some perfectly valid textual descriptions for two completely different products (a laptop and a battery) might differ in just one word: "*Acer TravelMate 4062LCI* with *battery*" and "*Acer TravelMate 4062LCI battery*".
2) **Short, undescriptive text:** Product offers typically come from merchants that expect to receive referrals from the online shopping engine. The e-commerce site typically has no control over the product description provided by the merchants, and in many cases the descriptions are brief and incomplete. For example, a product description from a merchant may just say '*P43A*', which is a model number of a motherboard. If the classifier is unaware of this model number, it will not be able to correctly classify the product.
3) **Discrepancy in vocabulary usage:** Offers are obtained from a very large number (typically in 1000s) of merchants. The merchants differ in their style, amount of specification, or the vocabulary in describing the products that they sell. Hence, any reasonable subset of offers that are manually labeled from this pool are too restrictive to capture all kinds of variability. Hence, there will be mismatch between the vocabulary

* Work done while author(s) were visiting MSR Search Labs

of words in the data set used to train the classifier, and in the offers to be categorized by the learned classifier. Figure 1 illustrates this mismatch between the term frequencies in product descriptions obtained from a dedicated group of aggregators, and from a large number of merchants. As an example, we can see from this plot that the term '*notebooks*' is more commonly used by merchants to refer interchangeably with the term '*laptop*' used predominantly by the aggregators.

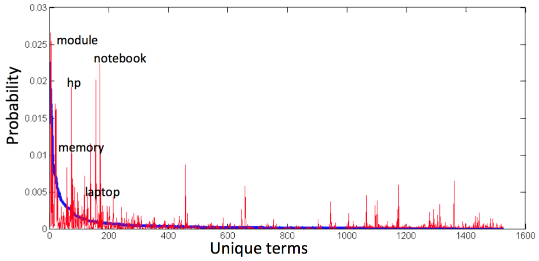

Figure 1. Differences in vocabulary between product descriptions from e-commerce aggregators (blue plot) and merchants (red plot, with spikes). Terms are sorted by the probability of occurrence in the descriptions from e-commerce aggregators.

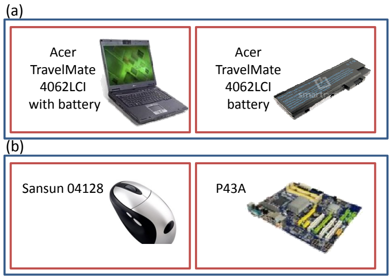
(a)
Acer TravelMate 4062LCI with battery
Acer TravelMate 4062LCI battery
(b)
Sansun 04128
P43A

Figure 2. Complexity of purely text-based classification in e-commerce setting: (a) Overlapping text across categories (b) Short undescriptive text

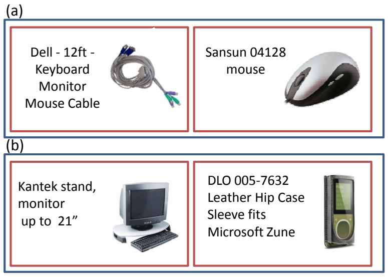
(a)
Dell - 12ft - Keyboard Monitor Mouse Cable
Sansun 04128 mouse
(b)
Kantek stand, monitor up to 21"
DLO 005-7632 Leather Hip Case Sleeve fits Microsoft Zune

Figure 3. Complexity of images: (a) Heterogenous categories (both offers belong to 'computing accessories' category) (b) Ambiguity of images

While incremental improvements can be obtained by trying new techniques that exploit textual features, in this paper we take a different approach. In particular, we observe that virtually all products in an e-commerce site have an associated image. This is because users shop visually, and they are thus more attracted to products with images. These images can be used in conjunction with text to improve classification. For example, consider Figure 2a, which shows a laptop and a battery. While their textual descriptions are exactly the same except for the word "with", their images are clearly different. In the same vein, in Figure 2b, even if the textual descriptions are uninformative, we can recognize from the images that they correspond to a computer mouse and a computer motherboard.

Using images for product classification presents its own challenges. Even after more than a decade of research in computer vision, image classification is a largely unsolved problem. Existing state-of-the-art image classifiers perform well only on certain categories and with limited amounts of variability [6]. In fact, a number of recent works in image classification have looked into using textual cues such as tags associated with similar images [8][17], and the textual content of webpages containing the images to improve classification performance [10]. In this paper, we ask the converse question:

Can images help improve text classification?

While some of the challenges posed in computer vision benchmark datasets (*e.g.,* PASCAL [6], Caltech 256 [7]) carry over in our application, there are also differences in our setting, that makes the problem even harder. Typically, classification tasks considered in the benchmark datasets focus exclusively on categorizing homogenous categories. In a product catalog, the categories can be quite heterogenous: a category in the product taxonomy often consists of products from very diverse sub-categories. As an example, in Figure 3(a), we observe that products (*e.g.,* mouse, power cords, *etc.*) from the same category ('Computer Accessories') can have completely different images. Moreover, the association between categories and images can be noisy. For example, in Figure 3(b), we observe an example of a product offer for a sleeve for a Microsoft Zune that shows the image of Zune itself as opposed to the sleeve. Because of these difficulties, an image-based classifier is significantly less accurate in our setting compared to a text-based classifier.

Not withstanding the above mentioned difficulties, in this paper we show that visual cues can be used to significantly improve classification accuracy. Although classification using generic multi-modal signals has been recently proposed in the literature, the use of images to aid text classification in a setting where there is significant difference in discriminative power between the two types of signals has not been systematically studied before. In fact, the only previous work [20] we are aware of is in the context of document classification. In that setting, several images in the document are used in conjunction with the text to identify the true category of a document. In our setting, an offer is provided with only a brief textual description and an image, and we would like to use the provided image as the additional signal to improve classification accuracy.

While the image signal can be very helpful as mentioned above, a few questions about its uniformity, processing cost, and general availability during application time remain. As with the text signal, image signal may also exhibit variability, especially when large number of data sources are involved, as in the current setting. In some cases, extracting image features during application time may be prohibitively expensive. While in some other situations, the image signal may be absent altogether. Given these challenges, an improved text-only classifier during application time is always desirable. So, we ask the question:

Can we learn an improved classifier based on textual features during *testing* time while exploiting text and image signals and unlabeled data available during *training* time?

In this paper, we make the following contributions:

- We initiate a study into the relatively unexplored setting where the image signals are less discriminative compared to text-based signals, and explore how image signals can be effectively used to improve the performance resulting from a purely text-based product classification.
- To address the challenges in product classification, we propose a novel classification algorithm: Confusion Driven Probabilistic Fusion++ (CDPF++). CDPF++ lessens the burden of image classifier needing to learn the entire decision surface by selectively learning multiple 3-way image classifiers that focus on category pairs on which the text classifier makes most errors.
- We performed all our experiments on real data obtained from the Bing Shopping catalog, demonstrating real-world significance of our proposed method, CDPF++. Our evaluation shows a 12% (absolute) improvement in precision at 100% coverage, and a 16% (absolute) improvement in recall at 90% precision compared to classifiers that only use textual description of products. Moreover, CDPF++ also results in a better text-only classifier, significantly outperforming the text-based baseline classifier.

The rest of the paper is organized as follows. In Section II, we provide an overview of the methods we use in this paper. In Section III, we present experimental results, and in Section IV, we discuss related work. We provide concluding remarks and thoughts for future work in Section V.

II. Classification by selectively combining text and image signals

In this section we outline the principles for selective combination of multi-modal signals in training a product classifier. We then propose our product classifiers following these principles.

A. Design principles for combining text and images

There is a large body of literature on learning classifiers that combine multiple views of the same signal source (*c.f.* [12] for a survey). Inspired by their success in various applications, these methods are also used in multimodal settings (*c.f.* [8]), where the source of the signals, or alternatively their modalities, are different. Examples of multiple modalities include speech waveforms and textual transcripts, or speech waveforms and images. It is often the case that different modalities vary in their discriminative power needed for various classification tasks. This is particularly true in our setting: textual cues arising from the free textual product description are more discriminative than that of image cues. In fact, in an experiment where we compared text-based classifier with image-based classifier, the performance of former was 74.5% while that of the latter was 54.7%, further validating that the image cues is weaker. In this setting, any method that combines text and image signals should have the following characteristics:

- **Cognizance of signals' discrimination capabilities**: The method should be aware of the disparities in discriminative abilities of various signals so that best improved classification performance is achieved when they are combined.
- **Use weaker signal selectively**: Instead of using the image signal (the weaker signal compared to text) to learn the entire discriminative surface over a large number of categories, it should be selectively used to learn a decision surface that discriminates a much fewer number of categories. This drastically reduces the complexity of the classification task involving image signals.
- **Ability to adapt**: The method should be able to adapt to changing vocabularies of textual descriptions, as described in Section I.
- **Improved Text-only classifier**: To handle situations where the image signal is expensive to obtain or is not present during application time, the method should be capable of building an improved text-only classifier by exploiting the image signal available during training time.

Based on the above principles, we propose a Confusion Driven Probabilistic Fusion (CDPF) approach to training a product classifier with labeled training data, which follows the first two of our design principles. We then present CDPF++, an extension of CDPF that uses a semisupervised learning approach to exploit the regularity of the unlabeled data to further improve the product classifier. CDPF++ follows all of our four design principles.

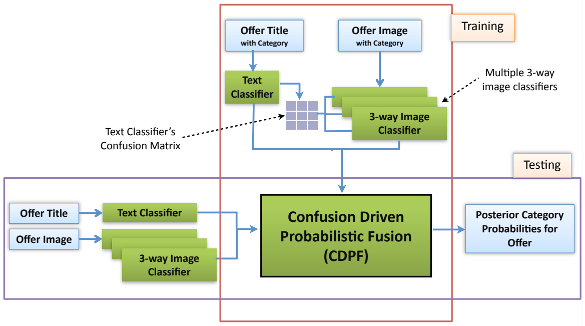

Figure 4. Confusion Driven Probabilistic Fusion (CDPF) uses multiple 3-way image classifiers to help reduce text classifier's confusions.

Algorithm 1 Confusion Driven Probabilistic Fusion (CDPF)

```
**Input**: $L = \{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^n, y^n)\}$ of $n$ labeled instances, where $\mathbf{x}_T^j$ and $\mathbf{x}_I^j$ are the text and image features of instance $\mathbf{x}^j$; $\eta$: the maximum number of image classifiers trained
**Output**: $H$, a classifier trained on $L$ exploiting both text and image features
Split $L$ into $B$ and $S$
$H_T = \textsc{TrainClassifier}(\{(\mathbf{x}_T, y) \mid (\mathbf{x}, y) \in B\})$
$C_T = \textsc{GetConfusionPairs}(H_T, \{(\mathbf{x}_T, y) \mid (\mathbf{x}, y) \in S\})$
/* Train a 3-way image classifier for each confusion pair, */
/* upto a maximum of $\eta$ such classifiers.                */
$H_{pool} = \emptyset$
**for all** $(c_1, c_2) \in \textsc{TopConfusionPairs}(C_T, \eta)$ **do**
    $F = \{(\mathbf{x}, y) \in S \mid y = c_1$ or $y = c_2\}$
    **for all** $(\mathbf{x}, y) \in S - F$ **do**
        $F = F \cup \{(\mathbf{x}, \perp)\}$
    **end for**
    $H_{c_1 c_2} = \textsc{TrainClassifier}(\{(\mathbf{x}_I, y) \mid (\mathbf{x}, y) \in F\})$
    $H_{pool} = H_{pool} \cup H_{c_1 c_2}$
**end for**
/* Use the text and all 3-way image classifiers to embed */
/* instances in a space of class membership probabilities.*/
Define $L'_s = \emptyset$
**for all** $(\mathbf{x}, y) \in S$ **do**
    $\mathbf{z}'_T = \textsc{GetPredictionProbabilities}(H_T, \mathbf{x}_T)$
    $\mathbf{z}'_I = \emptyset$
    **for all** $h \in H_{pool}$ **do**
        $\mathbf{z}_I^h = \textsc{GetPredictionProbabilities}(h, \mathbf{x}_I)$
        $\mathbf{z}'_I = \mathbf{z}'_I \cup \mathbf{z}_I^h$
    **end for**
    $\mathbf{z}' = (\mathbf{z}'_T, \mathbf{z}'_I)$
    $L'_s = L'_s \cup \{(\mathbf{z}', y)\}$
**end for**

$H = \textsc{TrainClassifier}(L'_s)$
```

B. Confusion Driven Probabilistic Fusion (CDPF)

We describe CDPF as it applies in our setting.

1) Training CDPF: Algorithm 1 describes the algorithm for training a CDPF classifier. We first identify those categories in which the the text classifier is most highly confused. For this, we divide the training data L into two non-overlapping subsets, B and S. We use B to train a text classifier, H_T, and also to identify the top η confusing pairs of categories, C_T (GETCONFUSIONPAIRS). For each of these pairs, we learn a 3-way image classifier between the confusing pairs, and an additional background category ($\perp$) that consists of all categories other than the pair under consideration. $\perp$ captures the possibility that the true category can be different from the categories in the pair. Note that the choice of η determines how much of the image information is being used.

Once all the $\eta + 1$ classifiers are trained, each $(\mathbf{x}, y) \in S$ is represented using $\mathbf{z}$ which consists of $\mathcal{K} + 3 \times \eta$ features, $\mathcal{K}$ of which are the prediction probabilities of H_T, while the remaining $3 \times \eta$ features are the prediction probabilities from the η 3-way image classifiers. These instances along with their labels are used to create a new training set L'_s. The final classifier H is then trained on this new labeled dataset.

To train various component classifiers (*e.g.,* H, H_T, *etc.*), CDPF assumes availability of a supervised classifier training method, TRAINCLASSIFIER. Any standard supervised classifier training algorithm, *e.g.,* Logistic Regression, Support Vector Machine (SVM), may be used as TRAINCLASSIFIER.

2) Discussion: CDPF builds on the concept of 'stacked generalization' [23] which provides a rich framework for combining varied feature sets and classifiers for increased robustness and generalization. An instantiation of 'stacked generalization' is Probabilistic Fusion (PF) wherein separate base classifiers are trained, independently for each signal and consequently the outputs from these classifiers, now in the same space of prediction probabilities, are combined to learn the final classifier.

Unlike in the case of Probabilistic Fusion, the use of multiple 3-way image classifiers creates easier classification task for the image classifier [9]. Since the 3-way image classifier focuses on the pairs of categories that are most confusing for the text classifier, we can improve on the overall classification task, especially when the two modalities are complementary to each other. This way, CDPF is cognizant of the discriminative capabilities of the signals, and selectively uses the weaker signal (image in our case), satisfying the first two requirements mentioned in Section II-A. In Section III, we present results that establishes complementarity of the two kinds of signals.

C. Leveraging unlabeled data: CDPF++

As we discussed in Section II-A, we would like the classifier to adapt to changing vocabularies in the specifications (*e.g.,* new products). In addition, we would like to obtain a text-based classifier that has successfully captured the information in the image signal. To achieve these two properties, we extended CDPF to CDPF++. The idea behind CDPF++

is that we make use of abundant unlabeled data to both (a) adapt to changing vocabulary without requiring manual labeling and (b) infuse the information in the image cues to the text-only component of the classifier. We capture these two properties by learning CDPF++ using a semisupervised technique based on the self-training paradigm [25].

CDPF++ is learned in an iterative fashion: Starting with CDPF, it is iteratively re-learned by making using of additional data that is automatically labeled (with high confidence) using the classifier trained in the current iteration. We also allow for the label of an automatically labeled instance to change in subsequent iterations, as it enables recovery from possible misclassification in the previous iterations.

Unlike in the traditional self training setup where the classifier that is re-trained is based on a single type of feature, CDPF++ makes use of both image and textual signals during learning. CDPF++ achieves all the requirements stated in Section II-A. In particular, it results in a better text-only component since information from the images are fed by adding automatically labeled data that used both image and text features. In addition, it adapts by using the unlabeled data.

III. Evaluation

In this section, we evaluate the following:

- **Value of images & unlabeled data:** Compared to a text-based classifier, is improved product classification performance possible by exploiting image signal & widely available unlabeled data? This is our main evaluation setting. (Section III-C)
- **Improvement of Text-only classifier:** Images may not always be available during classification time. For instance, when classifying millions of offers on a daily basis (common in commerce search setting), image processing may create a considerable overhead and hence may not be performed. We want to investigate how the text component of classifiers that make use of both signals (*e.g.,* CDPF++) can perform compared to a pure Text-based classifier that only uses textual features during training. (Section III-D)
- **Ablation Studies:** Can images (in the *absence* of unlabeled data) also help improve classification performance over text-based classifier? What is the effect of unlabeled data? (Section III-E)

We also perform qualitative analysis in Section III-F to study complementarity of the two kinds of signals, and also their learned roles in the classification task.

A. Algorithms used for Comparison

We compare the following algorithms:

- **Text & Text++**: *Text* is a classifier learned using only the textual features (and labels) of the labeled data. Its self-trained counterpart is Text++. Starting with Text, Text++ is iteratively re-learned using, both, labeled data and a portion of the unlabeled data (along with their predicted labels) that are confidently predicted by the classifier at the current iteration. Text is our main baseline algorithm.
- **Concat**:Concat is a classifier learned on features obtained by concatenating text and image features.
- **PF & PF++**: An instantiation of 'stacked generalization' is Probabilistic Fusion (PF), wherein separate base classifiers are trained, independently, for each signal and consequently the outputs from these classifiers, now in the same space of prediction probabilities, are combined to learn the final classifier. PF++ is the self-trained counterpart of PF that also uses unlabeled data along with their predictions to iteratively learn a better classifier.
- **CoTraining**: CoTraining [4] has been successfully used in a variety of multi-view problems in domains such as computer vision (*c.f.* [15], [21]) where there is abundance of unlabeled data and limited amount of labeled data. It is also an iterative algorithm in which separate classifiers are trained for each view, and the most confidently labeled instances from these separate classifiers are added to the training pool for re-training.
- **CDPF & CDPF++**: CDPF is described in Section II-B, while CDPF++ is presented in II-C. CDPF++ is the main algorithm proposed in this paper.

B. Experimental Setup

1) Dataset: The dataset used for the experiments in this section comprised of $\mathcal{K} = 17$ categories related to computing (*e.g.,* Laptop Computers, Motherboards, Computer Memory, *etc.*). This dataset was collected from the Bing Shopping catalog, a major Commerce Search engine. We used a total of N = 17989 instances for training and 10026 instances during testing. The test dataset is highly skewed in terms of number of instances per category, with average number of instances per category being 590. The largest and smallest categories have 1410 and 129 instances, respectively. We perform experiments in a transductive setting, and construct the unlabeled data by hiding labels from test instances.

The categories chosen for our experimentation overlap both in textual content (*e.g.,* vocabulary used in describing laptops and desktops, or cameras and camcorders), and also in image content (*e.g.,* a mouse from computing accessories category shown along with a desktop).

Additionally, we also experimented with data from 5 camera related categories (*e.g.,* Camcorders, Digital Cameras, Lenses, *etc.*). We observed similar trends as the results presented in this section and hence we omit the details due to lack of space.

2) Features used: For textual representation, we first derived lexical features based on standard tokenization of product descriptions into unigrams and bigrams. We used

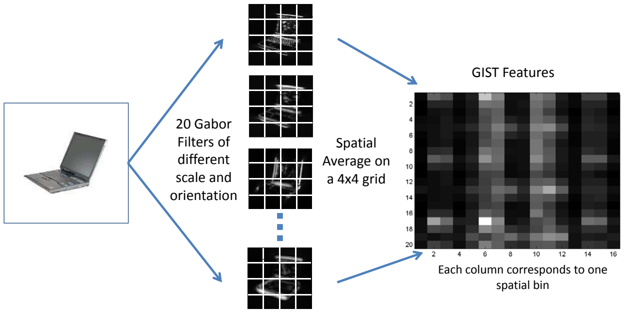
20 Gabor Filters of different scale and orientation
Spatial Average on a 4x4 grid
GIST Features
2
4
6
8
10
12
14
16
18
20
2
4
6
8
10
12
14
16
Each column corresponds to one spatial bin

Figure 5. Gist representation of an image.

boolean feature representing the presence or absence of these n-grams, as they tend to uniquely occur in offer descriptions.

For image representation, we capture the global structure of the objects in the image using a combination of textural property and their spatial layout (*c.f.* [22]). Textural property is captured, at each pixel in the image, using the responses of steerable pyramid tuned to 6 orientations and at 5 different scales. The spatial information is captured by dividing the image into 4×4 local grids, and computing the mean value of the magnitude of the response within those grids. The resulting $4{\times}4{\times}30$ vector is used as the feature representation of the image. Figure 5 shows an example image and its corresponding gist feature represented as an image.

3) Base Classifier: Methods described in this paper, including CDPF++, make use of a standard classifier as one of their components. We use Multiclass Logistic Regression for this purpose. This classifier is parameterized by $\mathbf{W}$, a collection of weight vectors $\{\mathbf{w}_k\}$, one for each category $k \in \mathcal{K}$. Each component w_{jk} measures the relative importance of the j^{th} feature for predicting k^{th} label. The logistic regression learns a mapping from the feature vector of instance $\mathbf{z}$ to the label y, using the following softmax logistic function:

$$P(y = k \mid \mathbf{z}, \mathbf{W}) = \frac{\exp\left(b_k + \mathbf{z} \cdot \mathbf{W}_k\right)}{1 + \sum_{j=1}^{\mathcal{K}} \exp\left(b_j + \mathbf{z} \cdot \mathbf{W}_j\right)} \quad (1)$$

where b_j $(1 \leq j \leq \mathcal{K})$ are bias terms. Given a labeled dataset $L = \{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^n, y^n)\}$, logistic regression learns the parameters $\mathbf{W}$ so as to maximize the conditional log-likelihood of the labeled data:

$$\mathbf{W}^* \quad \leftarrow \quad arg\,max_{\mathbf{W}} \sum_{j=1}^{n} \log P(y^j \mid \mathbf{x}^j, \mathbf{W}) \quad (2)$$

We call this classifier training process TRAINCLASSIFIER (see Algorithm 1). In Algorithm 1, the GETPREDICTIONPROBABILITIES$(H, \mathbf{z})$ method returns a $\mathcal{K}$-dimensional vector of category membership probabilities for instance $\mathbf{z}$, where each membership probability is calculated using Equation 1.

4) Performance metrics: For evaluation purposes, we have access to a test set of product offers, $u \in \mathcal{U}$. We also know the true category of these offers, $c_u^* \in \{1, \cdots, \mathcal{K}\}$, for every $u \in \mathcal{U}$. The classifier does not have any knowledge about the true category but predicts the best category $\tilde{c}$ with probabilistic score $\gamma_{u,\tilde{c}}$. By best category, we mean that there is no other c whose probability is higher than $\gamma_{u,\tilde{c}}$. We require the probabilistic score to be at least $\theta \in [0, 1]$ before calling it out as the correct category. Since the number of test examples vary largely across categories, we measure performance using micro average precision, recall and coverage at threshold level θ as:

$$\text{Precision}(\theta) = \frac{\sum_{u \in \mathcal{U}} I[(\gamma_{u,\tilde{c}} \geq \theta) \text{ AND } (c^* = \tilde{c})]}{\sum_{u \in \mathcal{U}} I[\gamma_{u,\tilde{c}} \geq \theta]} \quad (3)$$

$$\text{Recall}(\theta) = \frac{\sum_{u \in \mathcal{U}} I[(\gamma_{u,\tilde{c}} \geq \theta) \text{ AND } (c^* = \tilde{c})]}{|\mathcal{U}|} \quad (4)$$

$$\text{Coverage}(\theta) = \frac{\sum_{u \in \mathcal{U}} I[(\gamma_{u,\tilde{c}} \geq \theta)]}{|\mathcal{U}|} \quad (5)$$

where I[z] is the indicator function. For enabling rich e-commerce experience, precision of the classification task becomes highly important. Hence, we use metrics that compares precision of the system and also the recall at the desired precision level, to evaluate the effectiveness of the various algorithms:

1) **Precision at 100% coverage:** This is the overall precision of the algorithm, and computes the fraction of all our test offers ($\theta \geq 0$ in Equation. 5) that are correctly classified.
2) **Recall at a particular precision:** This measures the improvement in recall, for a particular desired precision level. We use 90% precision levels to compare improvements in recall. Note that, we choose different θ for each method in order to achieve the desired precision level.

5) Parameters used: For CDPF and CDPF++, we set $\eta = 50$. We found that the algorithm is not sensitive to this parameter suggesting that η can be set reliably without much tuning. For iterative re-learning in semi-supervised settings, we used 10 iterations as we found that to be sufficient for convergence. Also, for these setting, we only label instances when they are confidently predicted with a probability of at least 0.9.

We now return to the main evaluation questions presented at the beginning of Section III.

C. Value of images & unlabeled data

Figure 6 (our main result) compares the performances of various algorithms: *Text*++ uses unlabeled data, while CoTraining, PF++, and CDPF++ exploit both images and unlabeled data. Please note that algorithms also make use of text-based features. We make the following observations:

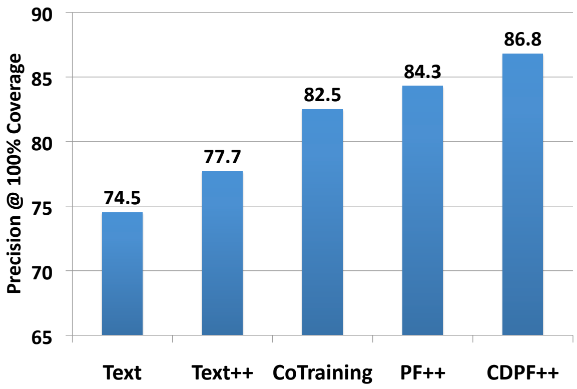

Figure 6. Image cues along with unlabeled data can help improve precision compared *Text*. CDPF++ is the proposed algorithm. This is our main result, please see Section III-C for more details.

1) *Text*++ shows improvement over *Text*. However, its performance is much lower than that of methods that utilize image cues, reinforcing the fact that the image signals provide information more and above that of the textual signals.
2) CoTraining shows better performance than the text self training (an improvement of 5% in classification accuracy), making the case for visual cues aiding in bootstrapping the correct category. This is further reinforced by a recent study [14] that showed that CoTraining can improve classification performance only when the signals provide additional information.
3) CoTraining performance (82%) is lower than that of CDPF++ and PF++. The reason for this is that during prediction time, cotraining does not leverage information from both signals as separate classifiers (based on text or image) are learned concurrently.

Method	Text	CoTraining	PF++	CDPF++
Recall	78.8	86.1	92.3	95.3

Table I
COMPARISON OF RECALL AT 90% PRECISION LEVEL

In Table I we also compare against *Text* the recall at 90% precision level of three algorithms that make use of both image signal and unlabeled data: CoTraining, PF++, and CDPF++. We can see that the recall of PF++ and CDPF++ are significantly better than that of purely text based classifier, supporting the usefulness of visual cues and unlabeled data in improving both precision and recall, simultaneously.

D. Improved Text-only Classifier

Here, we are interested in evaluating the performance of the text-only classifier component of the three methods, CoTraining, PF++, and CDPF++, compared to the purely text-based classifier, Text. An improved text-only classifier will be extremely useful in scenarios where image signals are not available during application time, due to factors such as image processing overhead.

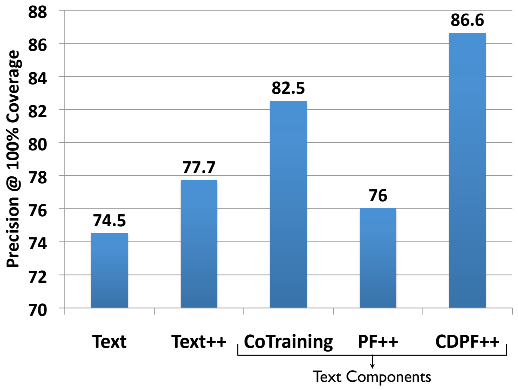

Figure 7. Comparison of text classifier components of various algorithms against *Text*. Note that the improvement in text components for CoTraining, PF++ and CDPF++ are due to use of image signals (see Section III-D).

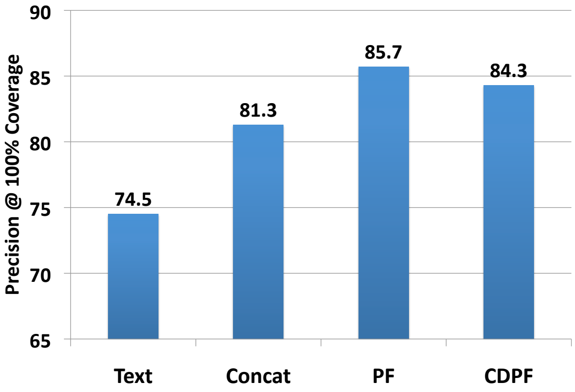

Figure 8. Using images (even in the absence of unlabeled data) helps improve precision @ 100% coverage.

Figure 7 compares the performance of the text classifier components of CoTraining, PF++, and CDPF++. We see that in the case of CDPF++, the text-only component classifier achieves similar performance compared to the full CDPF++ classifier that uses both signals (see Figure 6). This is because, during the self training process, information from the most confidently predicted unlabeled instances are teased out and factored into the text classifier. On the other hand, in the case of PF++ the unlabeled examples that are most confidently predicted are the ones that PF++'s text classifier were already confident of and hence the text-only component of the classifier did not improve its performance.

E. Ablation Studies

In this section, we study separately the value of images and of unlabeled data for the product classification task.

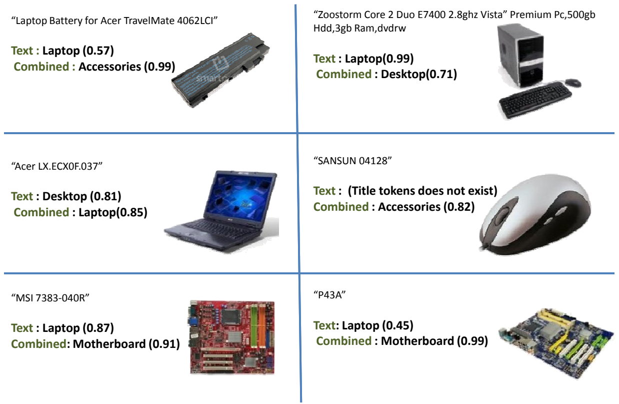
"Laptop Battery for Acer TravelMate 4062LCI"
Text : Laptop (0.57)
Combined : Accessories (0.99)
"Zoostorm Core 2 Duo E7400 2.8ghz Vista" Premium Pc,500gb Hdd,3gb Ram,dvdrw
Text : Laptop(0.99)
Combined : Desktop(0.71)
"Acer LX.ECX0F.037"
Text : Desktop (0.81)
Combined : Laptop(0.85)
"SANSUN 04128"
Text : (Title tokens does not exist)
Combined : Accessories (0.82)
"MSI 7383-040R"
Text : Laptop (0.87)
Combined: Motherboard (0.91)
"P43A"
Text: Laptop (0.45)
Combined : Motherboard (0.99)

Figure 9. Example offers showing correct classification after images were used in conjunction with text

1) Value of Images: Figure 8 compares *Text* with the methods that exploit both image and textual cues. We can see from the figure that classification accuracy has substantially improved (over 10%) by making use of visual features in addition to textual features. This improvement in performance establishes that images can help improve classification accuracy.

The difference in performance between Concat and CDPF/PF can be understood as follows: Concat learns interactions between the image and text signals in their feature space while CDPF and PF learns these interaction at a much reduced dimensional space, in the space of probabilistic outputs of classifiers trained independently on the image and textual features. This enables PF and CDPF to more succinctly capture the intrinsic discrimination capability of the feature types. Hence, PF and CDPF are more robust than Concat and is thus reflected in the higher gain in performance.

The drop in CDPF's performance compared to PF may be attributed to the fact that from the same (limited) amount of training data, CDPF has to estimate a larger number of parameters compared to PF. In the absence of appropriate regularization, this may lead to overfitting and thereby lower performance. However, with the availability of more training data, obtained by automatically labeling unlabeled data through self-training, this problem can be overcome, as in Figure 6 we observe that CDPF++ outperforms PF++.

In Figure 9, we present illustrative examples to showcase usefulness of image cues. For instance, a laptop battery offer which was classified by *Text* as belonging to 'laptop computers' category is correctly classified to accessories category. A 'mouse' offer that *Text* could not classify due to lack of features is classified correctly into the accessories category because the image of the offer helped to identify the category.

2) Value of Unlabeled Data: From Figures 6 and 8, we observe that CDPF++ outperforms CDPF (86.6% vs. 84.3%). The only difference between the two algorithms is the use of unlabeled data in case of CDPF++. This clearly demonstrates the additional benefit that unlabeled data can bring to the product classification task. Figure 10 shows examples to illustrate this. The images used in these two offers are atypical (in fact, ambiguous) representation of laptop and desktop, respectively. However, by corroborating multiple evidence from unlabeled data, CDPF++ corrected the mistake of CDPF.

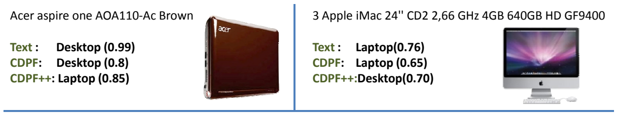

Figure 10. Example offers showing correct classification after using CDPF++

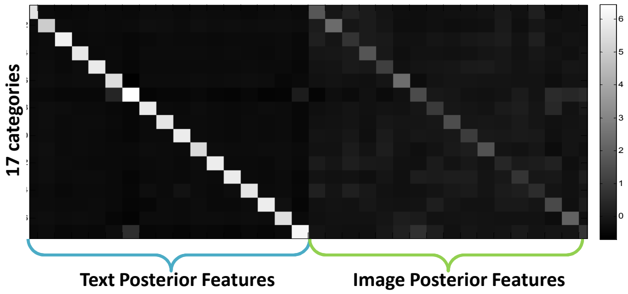

Figure 12. Second layer of learned weights of Probabilistic Fusion. Please see the figure on screen, for resolution and Section III-F2 for details

F. Qualitative Studies

1) Complementarity of text and image signals: We wanted to further understand if the reason for the improved performance using images is because image cues contain information different from that of the textual cues. To facilitate this, we learn two separate classifiers, one using text features and other using image features. Figure 11a-b shows the corresponding confusion matrices constructed using the predicted categories on a development set that was not used during training. The size of the black square at i^{th} row and j^{th} column correspond to the probability that instances whose correct category represented by the i^{th} row is labeled as j^{th} category. A perfect classifier is one in which there are no off-diagonal entries. We make the following observations:

1) Confusing categories for text classifier are often different from that for image classifier, indicating that they both carry unique bits of information
2) In categories where the text classifier is unreliable, image classifier dominates. As an example, while *Text* confuses 'Laptop computers' category with 'Desktop computers', image classifier can correctly discriminate this category.

The above observations convey the complementary nature of the two signals, and the reason for improved performance of classifiers that exploit image cues, in addition to textual cues.

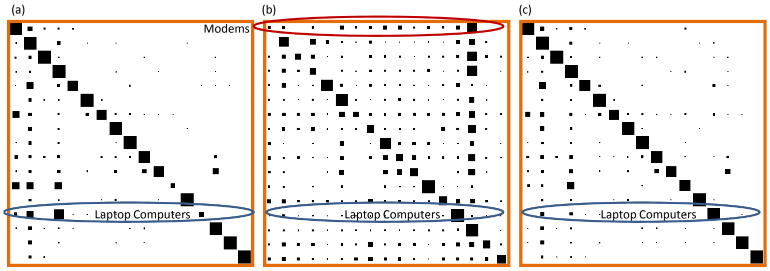

Figure 11. Comparison of confusion matrices of (a)Text based classifier (b) Image based classifier (c) CDPF

Figure 11c shows the confusion matrix of CDPF depicting that image cues enable improving over the text classifier's predictions.

2) Relative importance of text and image signals: We would like to understand how the classifiers make use of the image and text features. For this task, we study the PF classifier. Figure 12 shows second layer of learned weights using a 17 × 34 image corresponding to each of the 17 categories, and 34 features (17 each from text and image classifiers' predictions, respectively). As expected, since text feature is more dominant, predictions from the text classifier are weighted more than the image classifier's prediction (shown by brighter color on the text features). However, the non-negligible positive weights on the diagonal entries in image features indicate that the image component is also taken into account. Moreover, negative weights in the image component also indicates that the classifier learns to down weight the confidence in text component when there is non-supporting evidence from the image component.

We see similar trends in CDPF also, but for the sake of readability, we refrain from showing similar plots since the large number of 3-way image classifiers makes viewing cumbersome.

IV. Related Work

There has been a lot of recent interest in using multiple modalities to improve performance of classification tasks. In [16], classification of gene function is performed using both gene expression and phylogenetic data. In [5], speech and gesture classifiers are learned using audio visual data. In [24], a classifier is learned to infer if a person is more susceptible to Alzhemier's disease by combining two kinds of image data - MRI and tomography images.

There is also relevant work in computer vision where text associated with images are used for classification tasks. In [18], the authors classify broadcast news video by making use of closed text captions from news video, in addition to the visual features from image frames of the video. In particular, they looked at the binary classification problem of whether the news segment is about weather. [10] studies the problem of improving web image classification using contextual information in the form of web page content. Their approach is to learn separate classifiers for image and text and use hand coded fusion rules to combine the output of the two classifiers. In [17], tags associated with images are used in conjunction with image features to perform large scale landmark classification. In a more general image classification setting, [8] leverage textual tags in conjunction with image features.

In the above described applications, visual cues serve as the dominant signal, and text signal is used to complement the image features. In contrast, in our setting, the visual cues exhibit variability in terms of within category heterogeneity and in ambiguity across categories. Hence, we rely heavily on text information and use images (visual content) as a complementary signal to help improve the classification task.

To the best of our knowledge the only work that uses images to improve text classification was proposed in [20] where images were converted to a word using a joint classification and clustering procedure. However, the success of this algorithm depends on the availability of several images in a document as a single image document would be tantamount to a single word document which would be very unlikely to be suggestive of the true class. In our setting, in contrast, an offer is provided with only a brief textual description and an image, making this approach in [20] inapplicable.

There is a large body of work in combining classifiers (see [12] for a theoretical framework). These combination strategies are driven by the multiview setting where the goal is to make use of multiple feature subsets or different portions of the data to learn separate classifiers. Of special interest is the work of Ko *et.al.* [13] in which classifiers were combined together using confusion matrix constructed by analyzing a pair of classifiers, at a time. In CDPF we, instead, make use of confusion matrix of the classifier built using dominant signal (text) to construct three way classifiers of the other signal (image). It can also be interesting to see how [13] can be further used to combine all these three-way classifiers.

Much work in combining multi modal signals has continued to use this rich literature in multiview learning to combine signals from the different modalities (*c.f.* [18]). In this paper, we also investigated some of these techniques.

There is also a vast literature on semi-supervised learning techniques. The recent book of [25] provides a good reference. In this paper, we also studied two commonly used semisupervised methods, self training [25] and cotraining [4] that has shown to be useful in multiview and also multimodal setting *c.f.* [15], [21].

V. Summary and Future work

In this paper, we initiated a study into the relatively unexplored classification setting involving text and image signals, where the image signals are less discriminative compared to text-based signals, and explored how image signals can be used to complement text classifiers. We focus on the domain of product classification where textual product descriptions are brief and overlap in vocabulary across multiple categories. Further, due to the nature of application, the vocabulary of the product descriptions can also vary across merchants generating them.

To address these issues, we propose a novel algorithm Confusion Driven Probabilistic Fusion++ (CDPF++) which learns a number of three-way image classifiers focused only on those confusing categories of the text signal so as to capture the region of the discriminating surface that the dominant text classifier is unable to capture. Moreover, by exploiting unlabeled data, CDPF++ is able to adapt to changes in vocabulary. Through a variety of experiments on datasets from a major Commerce search engine's (Bing Shopping) catalog, we observed a 12% (absolute) improvement in CDPF++'s precision at 100% coverage compared to classifiers that only use textual description of products; and a 16% (absolute) improvement in recall at 90% precision over the same baseline. Moreover, CDPF++ also results in a better text-only classifier, significantly outperforming the text-based baseline classifier.

There are a number of interesting directions for future research including theoretical analysis of CDPF++, devising automatic schemes to decide on the number of image classifiers to train, and also apply CDPF++ to problems in other domains.

Acknowledgment

We would like to profusely thank Rakesh Agrawal for suggesting the problem, and for valuable discussions. We also thank Ariel Fuxman and Panayiotis Tsaparas for their valuable inputs.

References

[1] Forrester forecast: Online retail sales will grow to $250 billion by 2014. http://techcrunch.com/2010/03/08/forrester-forecast-online-retail-sales-will-grow-to-250-billion-by-2014, Accessed Feb 15 2011.

[2] Global trends in online shopping: A nielsen global consumer report. http:// hk.nielsen.com/ documents/ Q12010OnlineShoppingTrendsReport.pdf, June 2010.

[3] R. Agrawal and R. Srikant. On integrating catalogs. In *Proceedings of the 10th international conference on World Wide Web*, 2001.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.

[5] C. M. Christoudias, K. Saenko, L.-P. Morency, and T. Darrell. Co-adaptation of audio-visual speech and gesture classifiers. In *Proceedings of the 8th international conference on Multimodal interfaces*, 2006.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html.

[7] J. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[8] M. Guillaumin, J. Verbeek, C. Schmid, I. Lear, and L. Kuntzmann. Multimodal semi-supervised learning for image classification. In *IEEE Conference in Computer Vision and Pattern Recognition*, 2010.

[9] C. Hsu and C. Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

[10] P. R. Kalva, F. Enembreck, and A. L. Koerich. Web image classification based on the fusion of image and text classifiers. In *International Conference on Document Analysis and Recognition*, 2007.

[11] A. Kannan, I. Givoni, R. Agrawal, and A. Fuxman. Matching unstructured offers to structured product descriptions. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011.

[12] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *Pattern Analysis and Machine Intelligence*, 2002.

[13] A. H. R. Ko, R. Sabourin, A. Britto, Jr., and L. Oliveira. Pairwise fusion matrix for combining classifiers. *Pattern Recognition*, 40, 2007.

[14] M.-A. Krogel and T. Scheffer. Multi-relational learning, text mining, and semi-supervised learning for functional genomics. *Machine Learning*, 57:61–81, 2004.

[15] A. Levin, P. Viola, and Y. Freund. Unsupervised improvement of visual detectors using cotraining. In *IEEE International Conference on Computer Vision*, 2003.

[16] T. Li and M. Ogihara. Semi-supervised learning from different information sources. *Knowledge Information Systems Journal*, 7(3), 2005.

[17] Y. Li, D. J. Crandall, and D. P. Huttenlocher. Landmark classification in large-scale image collections. In *International Conference on COmputer Vision*, 2009.

[18] W. Lin and A. Hauptmann. News video classification using svm-based multimodal classifiers and combination strategies. In *In ACM Multimedia, Juan-les-Pins*, 2002.

[19] S. Sarawagi, S. Chakrabarti, and S. Godbole. Cross-training: learning probabilistic mappings between topics. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003.

[20] H. Shatkay, N. Chen, and D. Blostein. Integrating image data into biomedical text categorization. *Bioinformatics*, 22:446–453, 2006.

[21] F. Tang, S. Brennan, Q. Zhao, and H. Tao. Co-tracking using semi-supervised support vector machines. In *International Conference on Computer Vision*, 2007.

[22] A. Torralba, K. P. Murphy, W. T. Freeman, and M. Rubin. Context-based vision system for place and object recognition. In *International Conference on Computer Vision*, 2003.

[23] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[24] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen. Multimodal classification of alzheimers disease and mild cognitive impairment. *NeuroImage*, 2011.

[25] X. Zhu and A. B. Goldberg. *Introduction to Semi-Supervised Learning*. Morgan and Claypool, 2009.